Igoshina Ekaterina Dmitrievna
Student
Ural Federal University named after the first
President of Russia B.N. Yeltsin
Russia, Ekaterinburg
Academic supervisor: Gubina Dilyara Ilshatovna

EFFICIENCY OF THE ALGORITHMS FOR FORECASTING THE PROPERTIES OF MATERIALS BY THEIR MOLECULAR COMPOSITION

Abstract. *Numerous machine learning (ML) algorithms in materials science allow predicting the properties of materials by molecular composition based on known data. However, there are factors that impede the effective application of algorithms for solving a specific problem. This article discusses the main methods for predicting the properties of materials using ML algorithms and the level of their efficiency, identifies factors that affect the success of predictions, and put forward proposals for improving some algorithms.*

Keywords: *machine learning, material property prediction, quantitative structure-property relationship (QSPR), descriptor, combination of algorithms.*

Игошина Екатерина Дмитриевна
Студент
Уральский федеральный университет имени первого
Президента России Б.Н. Ельцина
Россия, г. Екатеринбург
Научный руководитель: Губина Диляра Ильшатовна

ЭФФЕКТИВНОСТЬ АЛГОРИТМОВ ПРОГНОЗИРОВАНИЯ СВОЙСТВ МАТЕРИАЛОВ ПО ИХ МОЛЕКУЛЯРНОМУ СОСТАВУ

Аннотация. *Множество алгоритмов машинного обучения (ML) в материаловедении позволяет на основе известных данных прогнозировать свойства материалов по молекулярному составу. Однако существуют факторы, препятствующие эффективному применению алгоритмов для решения конкретной задачи. В данной статье рассмотрены основные способы прогнозирования свойств материалов с помощью алгоритмов ML и уровень их эффективности, установлены факторы, влияющие на успешность предсказаний и выдвинуты предложения по улучшению некоторых алгоритмов..*

Ключевые слова: *машинное обучение, прогнозирование свойств материалов, количественное соотношение структура-свойство (QSPR), дескриптор, сочетание алгоритмов).*

The introduction of machine learning into the field of modern materials science has contributed to the emergence of a huge number of algorithms, methods and systems that allow you to classify and predict data. A special place among them is occupied by the automatic prediction of the properties of materials according to the given parameters of the chemical composition. The use of these methods can significantly reduce the expenditure of funds and materials for research, facilitate the processing of information on the properties of materials that are difficult to measure or calculate using traditional methods - because of monetary, time or other difficulties [1, p. 1]. Machine learning algorithms solve the problem of studying the correlation between individual properties, creating models for predicting quantitative relationships of both individual properties and their structures. Such methods are used to solve specific analytical problems within the framework of a common project and require significant external resources. The effectiveness of machine learning algorithms directly depends on many factors, such as the availability of free data on the properties and structure of materials, the sample size, the influence of third-party parameters on changing the properties and internal structure of the material. This article will consider the main methods of predicting the properties of materials using machine learning algorithms, the level of efficiency of these methods when changing external conditions and

sampling and identifies factors that affect the degree of success in predicting the qualities of materials.

For a long period, experimental observations and finding correlations between various parameters have been the main means of studying and understanding the various chemical and physical properties of materials. The properties of materials, such as hardness, melting point, glass transition, solidification, ionic conductivity, molecular sputtering energy and lattice constant, are described both at the micro- and macroscopic levels [10, p. 162]. Such versatile studies have formed many expert databases that accumulate the results of many years of work to identify signs and dependencies between the parameters of materials. The demand for such platforms as Materials Project, Citrination, Materials Data Facility, Aflowlib and OQMD [3] is determined by the versatility and relative availability of data. Data sets from such large data warehouses become difficult to interpret and analyze using the old approaches. Data mining techniques allow you to create more accurate automated models for predicting individual properties of materials and their aggregates based on training on ready-made sets. An important role in the formation of a new approach to data analysis was played by the emergence of software libraries such as Matminer, Keras, TensorFlow [4]. Multifunctionality of the libraries, as well as their tight integration with skitit-learn for Python, makes it easier to work with large amounts of data and provides access to many useful utilities. Computational material design is expected to lead to the discovery of new materials and reduce the time and cost of material development. This is especially in demand for determining the direction of development of an enterprise when finding a way to design and produce the necessary materials with specified properties.

Machine learning methods, regardless of the ultimate goal of the study, use a finite known dataset as the basis. This is also the limitation of such algorithms. Algorithms of varying degrees of complexity require special test data corresponding to the essence of the task.

The formation of separate groups of parameters from a dataset is an important step in building a forecasting model. Since, often, it is sufficient to use a limited

number of parameters fixing the properties of materials, establishing dependencies between them becomes a priority task. Traditional compressed probing methods such as LASSO (Linear Absolute Compression and Selection Operator) and algorithms based on it are poorly applicable for cases where functions are correlated. The article [5] considers an improved systematic approach to the detection of descriptors (sets of parameters) based on the SISSO (sure-independence screening and sparsifying operator) method. This approach is a modification of LASSO and allows you to analyze huge function spaces, including correlated ones, find the optimal composition of the descriptor, and generate predictive models in the form of analytical formulas at the output. The optimal solution is distinguished from the combinations of functions related to the required material properties, even based on relatively small training sets. The SISSO method was used to predict the level of thermal equilibrium of chemical compounds by finding the best descriptor and was confirmed by experimental data [6]. At the same time, the quality of prediction allows us to assert that the algorithm is efficient. The use of the analytically identified descriptor made it possible to predict the Gibbs energy parameters with high accuracy for any structure that includes the elements of the descriptor (in particular, we are talking about the presence of the total DFT energy). This makes it possible to predict temperature-dependent thermodynamics with high throughput over a wide range of compositions and temperatures. The idea of expanding the functionality of the method to the multitasking level MT-SISSO was developed in the study [7]. This approach is especially suitable for databases of dissimilar materials with limited or partial data, for example, where not all properties are specified for all materials in the training set.

The SISSO model is one of the most general approaches to handling large datasets. The demand for general analysis is often much higher than narrowly focused algorithms. However, they take place and are based on the manual formation of optimal descriptors that do not capture a large set of features. In particular, this is relevant for predicting the energy of the forbidden zone or the ability to glass formation, although there have been literary descriptions and ready-made sets of parameters and dependences of these properties for many years [2, p. 1-2]. Such algorithms are used

to solve specific problems with a small spread of features, which significantly reduces their prevalence. Concentration on several indicators allows narrowly focused algorithms to have the highest performance and prediction accuracy in the selected area [8, p. 1]. However, even for special algorithms, there are signs that are difficult and ineffective to predict using computer programs, as is acceptable for the glass transition temperature of the material. This is due to multiple external and internal factors - pressure, molecular structure, conformational features [10, p. 162]. The problem of predicting such properties was solved in the study [17]. The solution consists in selecting a successful combination of several methods, generalized formation of a set of attributes, grouping the sample into chemically similar subsets, and then training each subsample separately from the rest.

Each method has individual advantages in speed, interpretability, accuracy, and data coverage. Parallel evaluation of the effectiveness of several algorithms allows you to establish the degree of utility of the algorithm with its computational requirements and the level of errors issued. Machine analysis approaches allow finding a balance between the real correspondence of predictions to experimental data and a low level of predicted errors. However, not for any algorithm such a balance is possible, which generates a lot of discussions and studies on this topic. So, for example, the degree of learning of decision tree ensembles and high classification accuracy are much higher than any other combinations. With this undeniable advantage, the results of this algorithm are difficult for a specialist to interpret, which complicates its full-fledged wide application in predicting the properties of materials [2, p. 2]. The advantages and disadvantages of using the trained tree algorithm for analyzing large amounts of data were previously discussed by us in a related study [13]. In particular, in a simplified form, this algorithm does an excellent job of handling data outliers.

A promising extension of the decision tree algorithm is the random forest regression model. The construction of the forecast is based on the generalization of the results of calculating several trees with different parameters and depends on the number of decision trees, the maximum depth of the trees and the number of random

features. Optimization of the main parameters is the main task when building a forecasting model. After creating training sets using the Bootstrap method, it is recommended to use the classification and regression tree (CART) method to select the best mode for splitting the set. Subsequently, the predicted value is determined by averaging the predicted values by the trees. This method was first investigated by Leo Breiman together with Adele Cutler [14]. In their research, they established the main advantages of the method, which make it popular and relevant for solving forecasting problems in materials science. The effectiveness of this forecasting technology was experimentally confirmed by research [9] to determine the porosity readings of ceramic materials. According to the results obtained, the porosity of the material predicted by the regression random forest model is within the acceptable range of measurement errors and directly depends on the input parameters of the algorithm.

The approach that involves finding the relationship between molecular features and macroscopic properties is called quantitative structure-property ratio (QSPR). The QSPR methodology is used in various studies and is applied to predict the properties of materials, such as: flash point [19], normal boiling point [20], Henry's law constants [21] and many others. To predict the properties of materials, as a rule, regression analysis methods are used that show the best results. The QSPR model for feature selection can use a wide range of algorithms, such as genetic algorithms (GA) [22, p. 53-59], stepwise regression (SR) [22, p. 55-59], a simple replacement method (RM) [22, p. 55-59] or enhanced replacement method (ERM). According to the source [23], it is the use of the ERM method that provides the best prediction based on a combination of regression methods and genetic algorithms. This method was used to identify the minimum number of possible descriptors before using the genetic programming (GP) method in the study [18, p. 5-16]. The same method was used as the basis for the formation of an improved model for predicting individual properties of crude oil in this study [24].

Most methods for predicting material properties at the macroscopic level are subject to multiple changes. Such modifications eliminate the disadvantages of the combined algorithms by overlapping them with each other. For example, based on

support vector regression (SVR), a hybrid methodology was created that combines genetic algorithms and SVR to predict atmospheric corrosion parameters for zinc and steel [15]. The research results prove the best predictive ability of the combination of algorithms. Subsequently, the poor performance of the component parts was eliminated by the introduction of two-stage SVR prediction based on the choice of characteristics (FSTS-SVR). The algorithm was analyzed by the authors of the study [10, p. 166-167] and showed the smallest percentage of prediction errors in comparison with previous versions.

Algorithms based on graph representations also show excellent performance when combined. Among them, the crystal graph convolutional neural network (CGCNN) or the material graph network (MEGNet) stands out [25]. This method became the basis for the formation of the optimal material descriptor network (MODNet) algorithm using some SISSO elements. In their research, the authors compare the developed MODNet method with the functionality of the methods included in it [8, p. 3-5]. The results of the analysis showed the superiority of this algorithm in most of the criteria, which indicates a successful experience in combining algorithms. A group of researchers [11, 12], based on individual internal research, concluded that such an iterative combination of teaching methods will allow achieving a high rate of discovery of both simple and complex compounds.

The microscopic characteristics of a material (atomic characteristics) are the basis of the macroscopic properties. The problem of predicting microscopic properties is the lack of a description of a wide range of parameters, instead of focusing on specific aspects (lattice structure, energy of bands and molecules) [10, p. 164]. However, with optimized input parameters, macroscopic prediction algorithms can show much more successful results than narrowly targeted algorithms. In the study [16], the already mentioned SVR method, generalized regression neural networks (GRNN), ANN, random forests and multiple linear regression were used to predict the lattice constants of complex cubic perovskites. At the same time, according to the experimental data of the study, the best prediction model was built by the SVR method, which indicates the undoubted advantage of the algorithm in data generalization [10, p. 166-168; 16].

Machine learning methods have tremendous potential in materials science. Their use in discovering new materials and predicting their properties by solving problems of classification, regression, probability assessment, sorting and selection of data is becoming the main direction of development of the region as a whole. However, it should be noted that no single data mining algorithm can achieve absolutely successful predictions for all properties and attributes of a material. In this regard, comparing the effectiveness of a variety of machine learning methods is a paramount task at the initial stage of model design. Data cleaning and the formation of a complex of descriptors is carried out using specialized methods (SISSO, MODNet, CART). The method of direct data analysis is also selected based on factors such as: the sample size, the shape of the cleaned data (in particular, the presence of categorical attributes), the ability to handle exceptions and outliers (classification algorithms do a good job with this), smoothness of the results, speed and performance of the algorithm. The use of one method or another should be determined by the ease of use, the openness of the functions of the algorithm, the ability to adapt the method to the conditions of the problem (including focusing on the individual characteristics of the material property) and the balance between the learning rate and the rate of real prediction. A theoretical study has shown the superiority of regression analysis methods and a combination of various algorithms over single methods. The overall efficiency of the algorithm when used to solve a specific problem in the field of materials science depends on the availability of optimal external and internal parameters.

REFERENCES

1. Ramprasad R. Machine learning in materials informatics: recent applications and prospects / R. Ramprasad, R. Batra, G. Pilania et al. // Computational Materials. – 2017. – Vol. 3, №54. – Text: electronic. – URL: https://www.nature.com/articles/s41524-017-0056-5 (Reference date 11.11.2020).

2. Ward L. A general-purpose machine learning framework for predicting properties of inorganic materials / L. Ward, A. Agrawal, A. Choudhary and C.

Wolverton// Computational Materials. – 2016. – Vol. 2, 16028. – Text: electronic. – URL: https://www.nature.com/articles/npjcompumats201628 (Reference date 11.11.2020).

3. Jha D. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning / D. Jha, K. Choudhary, F. Tavazza, W. Liao et al. // Nature Communications. – 2019. – Vol. 10, 5316. – P. 1-12. – Text: electronic. – URL: https://www.nature.com/articles/s41467-019-13297-w (Reference date 09.11.2020).

4. Ward L. Matminer: An open source toolkit for materials data mining / L. Ward, A. Dunn, A. Faghaninia, Nils E.R. Zimmermann // Computational Materials Science. – 2018. – Vol. 152. – P. 60-69. – Text: electronic. – URL: https://www.sciencedirect.com/science/article/pii/S0927025618303252 (Reference date 15.11.2020).

5. Ouyang R. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates / R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler et al. // Physical Review Materials. – 2018. – Vol. 2, 083802. – Text: electronic. – URL: https://arxiv.org/abs/1710.03319 (Reference date 13.11.2020).

6. Bartel C.J. Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry / C. J. Bartel, S. L. Millican A. M. Demi et al. // Nature Communications. – 2018. – Vol. 9, №4168. – Text: electronic. – URL: https://www.nature.com/articles/s41467-018-06682-4 (Reference date 14.11.2020).

7. Ouyang R. Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO / R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler et al. // Journal of Physics: Materials. – 2019. – Vol. 2, №2. – Text: electronic. – URL: https://iopscience.iop.org/article/10.1088/2515-7639/ab077b (Reference date 13.11.2020).

8. Breuck P. P. Machine learning materials properties for small datasets / P. P. De Breuck, G. Hautier, G. M. Rignanese// Materials Science (cond-mat.mtrl-sci). –

2020. – Text: electronic. – URL: https://arxiv.org/pdf/2004.14766v2.pdf (Reference date 13.11.2020).

9. Gao X. Porosity Prediction of Ceramic Matrix Composites Based on Random Forest / X. Gao, L. Wang, L. Yao // IOP Conference Series: Materials Science and Engineering. – 2020. – Vol. 768, 052115. – Text: electronic. – URL: https://iopscience.iop.org/article/10.1088/1757-899X/768/5/052115/meta (Reference date 15.11.2020).

10. Lui Y. Materials discovery and design using machine learning / Y. Liu, T. Zhao, W. Ju, S. Shi // Journal of Materiomics. – 2017. – Vol. 3. – P. 159-177. – Text: electronic. – URL: https://www.sciencedirect.com/science/article/pii/S2352847817300515 (Reference date 16.11.2020).

11. Meredig B. Combinatorial screening for new materials in unconstrained composition space with machine learning / B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal et al. // Physical Review B. – 2014. – Vol. 89, №094104. – Text: electronic. – URL: https://journals.aps.org/prb/abstract/10.1103/PhysRevB.89.094104 (Reference date 13.11.2020).

12. Hautier G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory / G. Hautier, C.C. Fischer, A. Jain, T. Mueller et al. // Chem. Mater. – 2010. – P. 3762–3767. – Text: electronic. – URL: https://pubs.acs.org/doi/10.1021/cm100795d (Reference date 16.11.2020).

13. E. Igoshina. Evaluation of effectiveness of algorithms for detection of suspicious bank transactions // International research to practice conference for educators, postgraduates and students. – Languages in professional communication. – 2020. – P. 522-527. – Text: electronic. – URL: http://hdl.handle.net/10995/84232 (Reference date 15.11.2020).

14. Kupenova E. M. Random forest method for satellite imagery classification problems / E. M. Kupenova, A. V. Kashnitsky. // TvSU Bulletin. Series "Geography and Geoecology". – 2018. – Vol. 3. – P. 99-107. – Text: electronic. – URL: https://elibrary.ru/item.asp?id=36815855 (Reference date 13.11.2020).

15. Fang S. F. Hybrid genetic algorithms and support vector regression in forecasting atmospheric corrosion of metallic materials / S. F. Fang, M. P. Wang, W. H. Qi, F. Zheng // Computational Materials Science. – 2008. – Vol. 44, Issue 2. – P. 647-655. – Text: electronic. – URL: https://www.sciencedirect.com/science/article/pii/S0927025608002371 (Reference date 17.11.2020).

16. Majid A. Lattice constant prediction of cubic and monoclinic perovskites using neural networks and support vector regression / A. Majid, A. Khan, G. Javed, A. M. Mirza // Computational Materials Science. – 2010. – Vol. 50, Issue 2. – P. 363-372. – Text: electronic. – URL: https://www.sciencedirect.com/science/article/pii/S0927025610004970 (Reference date 15.11.2020).

17. Ward L. A general-purpose machine learning framework for predicting properties of inorganic materials / L. Ward, A. Agrawal, A. Choudhary, C. Wolverton // Computational Materials. – 2016. – Vol. 2, №16028. – Text: electronic. – URL: https://www.nature.com/articles/npjcompumats201628 (Reference date 12.11.2020).

18. Ghomisheh Z. Prediction of critical properties of sulfur-containing compounds: New QSPR models / Z. Ghomisheh, A. E. Gorji, M. A. Sobati // Journal of Molecular Graphics and Modelling. – 2020. – Vol. 101, №107700. – Text: electronic. – URL: https://www.sciencedirect.com/science/article/pii/S1093326320304897 (Reference date 17.11.2020).

19. Torabian E. New structure-based models for the prediction of flash point of multi-component organic mixtures / E. Torabian, M.A. Sobati // Thermochimica Acta. – 2019. – Vol. 672. – P. 162-172. – Text: electronic. – URL: https://www.sciencedirect.com/science/article/pii/S0040603118303678 (Reference date 15.11.2020).

20. Abooali D. Novel method for prediction of normal boiling point and enthalpy of vaporization at normal boiling point of pure refrigerants: a QSPR approach / D. Abooali, M.A. Sobati // International Journal of Refrigeration. – 2014. – Vol. 40. – P.

282-293. – Text: electronic. – URL: https://www.sciencedirect.com/science/article/pii/S0140700713003861 (Reference date 16.11.2020).

21. Ghaslani D. Descriptive and predictive models for Henry's law constant of CO2 in ionic liquids: a QSPR study / D. Ghaslani, Z.E. Gorji, A.E. Gorji, S. Riahi // Chemical Engineering Research and Design. – 2017. – Vol. 120. – P. 15-25. – Text: electronic. – URL: https://www.sciencedirect.com/science/article/pii/S0263876217300011 (Reference date 17.11.2020).

22. Goodarzi M. Application of quantitative structure-property relationship analysis to estimate the vapor pressure of pesticides / M. Goodarzi, L. S. Coelho, B. Honarparvar, et.al // Ecotoxicology and Environmental Safety. – 2016. – Vol. 128. – P. 52-60. – Text: electronic. – URL: https://www.sciencedirect.com/science/article/pii/S0147651316300203(Reference date 17.11.2020).

23. Mercader A.G. Advances in the replacement and enhanced replacement method in QSAR and QSPR theories / A.G. Mercader, P.R. Duchowicz, F.M. Fernandez, E.A. Castro // Journal of Chemical Information and Modeling. – 2011. – Vol. 51(7). – P. 1575-1581. – Text: electronic. – URL: https://pubs.acs.org/doi/10.1021/ci200079b (Reference date 16.11.2020).

24. Abooali D. A new empirical model for estimation of crude oil/brine interfacial tension using genetic programming approach / D. Abooali, M.A. Sobati, S. Shahhosseini, M. Assareh // Journal of Petroleum Science and Engineering. – 2019. – Vol. 173. – P. 187-196. – Text: electronic. – URL: https://www.sciencedirect.com/science/article/pii/S0920410518308283 (Reference date 19.11.2020).

25. Xie T. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties / T. Xie, J. C. Grossman // Physical Review Letters. – 2018. – Vol. 120(14), 145301. – Text: electronic. – URL:

https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.120.145301 (Reference date 14.11.2020).